Proceedings of CONF-FMCE 2025 Symposium: Semantic Communication for Media Compression and Transmission
DOI: 10.54254/2755-2721/155/2025.GL23384

Analysis of Deep Learning Frameworks for Occluded Face Recognition

Zhihao Yan

College of Electronic Science and Engineering, Jilin University, Jilin, China
yanzh1922@mails.jlu.edu.cn

Abstract: Face recognition has been fully applied in more and more fields, such as security monitoring, financial services, social media and traffic management. However, masks, glasses, and other items often block the face, which is also affected by changes in light and posture. Therefore, efficiently dealing with face recognition based on occlusion is very important. With the development of deep learning, many neural network frameworks have been effectively applied to this problem. This paper summarises the process of occluded face recognition (OFR) and introduces some standard face recognition datasets. Convolutional Neural Networks (CNN), Variational Autoencoder (VAE) and Generative Adversarial Networks (GAN), which are three deep learning frameworks used for OFR, are demonstrated, and the last two of them are compared. A hybrid model of VAE and GAN is proposed. This model can avoid the common pattern breakdown problem of GAN and improve the stability of the training process. At the same time, the model can generate multiple possible occlusion cases and more realistic images, effectively improving the ability to process occlusion.

Keywords: Occluded Face Recognition, Convolutional Neural Networks (CNN), Variational Autoencoder (VAE), Generative Adversarial Networks (GAN)

1. Introduction

Facial recognition technology is a biometric identification methodology that utilises the unique physiological characteristics of human facial features to perform automated identity verification through pattern recognition and analysis [1]. Current identity verification methods encompass biometric technologies, including fingerprint recognition, iris scanning, and Deoxyribonucleic Acid analysis. However, facial recognition demonstrates superior reliability, security, and enhanced social acceptability compared to other biometric modalities, exhibiting substantial application potential [2]. While facial recognition constitutes an effortless cognitive process for humans, it presents significant technical complexity for artificial systems. When confronted with static occlusion (e.g., face masks, eyewear) and dynamic occlusion (such as postural variations and illumination fluctuations from environmental factors), facial recognition systems always fail to meet anticipated performance criteria under such conditions [3].

Prevalent occlusion removal methods, such as generative adversarial networks (GANs), need substantial computational overhead [4]. To address this problem, Liu et al. proposed an adaptive multi-type occlusion face recognition model (AMOFR) to address the challenges of face recognition under different occlusion conditions. It can effectively use the feature information under the occlusion area to solve the face recognition challenge under different occlusion conditions. The experimental

DOI: 10.54254/2755-2721/155/2025.GL23384

results show that AMOFR improves accuracy in treating multiple occlusions, such as glasses, sunglasses and masks [5]. Xu et al. proposed a face recognition algorithm based on a cyclic generative adversarial network (CycleGAN) to solve the impact of face image occlusion on recognition accuracy. By cyclic training of two generators and a discriminator, the method realizes blind restoration of blocked face images without original information. After the repair, the ResNet-50 network is used for face recognition, and the RegularFace loss function is introduced to improve the recognition effect. The experimental results show that compared with the traditional method, the recognition accuracy of the proposed algorithm is significantly improved under different occlusion types and areas. Especially when the occluded area is 40%, the accuracy rate is improved by 14.4 percentage points [6]. Since et al. propose a lightweight model for removing occlusion from face images, specifically using Pix2pix GAN technology [7]. The researchers significantly reduced the model size from 449MB to 117MB using dynamic quantisation while maintaining high image quality, with a Structure Similarity Index Measure (SSIM) of 0.896 and a Peak Signal-to-Noise Ratio (PSNR) of 27.32 db, superior to other existing methods. The model is suitable for running on resource-constrained devices, such as mobile devices and Internet of Things (IoT) devices, enabling real-time applications. The study shows that the quantified Pix2pix model is excellent at removing occlusion with low latency and L1 loss, demonstrating its potential for practical applications.

Recent years have witnessed substantial breakthroughs in theoretical advancements and practical applications of deep learning technologies. Therefore, more and more scholars have begun to use deep learning to solve the occlusion problem in face recognition. The primary purpose of this study is to exhaustively analyse the whole process of occluded face recognition (OFR) and the concepts of Convolutional Neural Networks (CNNS), Variational Autoencoder (VAE) and GAN and how they are used in OFR. The paper will systematically summarise the characteristics of standard face recognition datasets and compare VAE and GAN. Finally, a hybrid model is proposed, and its advantages in handling occlusion are analyzed.

2. Methodology

2.1. Datasets description

The advancement of facial recognition technology is highly dependent on the availability of datasets encompassing comprehensive demographic diversity and environmental variability. The following three datasets constitute the most extensively utilised benchmarks in academic research and industrial applications, addressing model training, performance benchmarking and validation under extreme operational conditions [1].

As the "benchmark gold standard" in face recognition, The Labelled Faces in the Wild (LFW) contains 13,233 network images in an unconstrained environment, covering 5749 people, focusing on challenging lighting, attitude, occlusion and other factors. Its data comes from public scenes such as news and sports events, and is labelled as a binary classification label of "whether the same person". It is dedicated to face verification (1:1 matching) evaluation, where all mainstream algorithms (e.g., DeepFace, FaceNet) must report recognition accuracy to demonstrate validity.

MS-Celeb-1M is a massive dataset (1 million celebrities, about 10 million images) published by Microsoft. It is characterised by high noise (which needs to be cleaned) and high complexity. The main body covers celebrities in various fields, such as entertainment and politics. It is mainly used for training deep face recognition models (such as ArcFace and CosFace).

Intelligence Advanced Research Projects Agency Janus Benchmark C (IARPA IJB-C), the extreme Scenarios dataset published by the Intelligence Advanced Research Projects Agency (IARPA), contains 3,531 people and 21,294 images and video frames, including cross-modal (still images and video), extreme lighting, heavy occlusion, and low-quality samples. It is designed for

DOI: 10.54254/2755-2721/155/2025.GL23384

high-risk scenarios such as military and security and promotes the study of difficult problems such as occlusion and motion blur, such as Partial Face recognition technology.

2.2. Methods

2.2.1. Introduction of occluded face recognition

The OFR process can be systematically summarized into a robust, multi-stage collaborative framework. As shown in Figure 1, the first step is preprocessing and feature localisation. A robust detector was used to locate the face area, and the pose differences were eliminated through key point alignment. At the same time, the occluded area is identified by the attention mechanism or segmentation network to provide prior information for subsequent processing. The second step is robust feature extraction, which uses a pre-trained depth model to extract global features. Local feature expression is enhanced by adaptive pooling or feature pyramid (FPN). The third step is occlusion compensation and recovery by GAN or VAE to reconstruct the obscured area and generate a complete face image. Feature interpolation or sparse representation is used to repair the occluded area directly in the feature domain. The fourth step is multimodal fusion. Integrating texture, geometric features and contextual information (such as clothing and hairstyle) in the visible area improves recognition robustness. The last step is confidence calibration, which estimates the confidence of the recognition results based on the area and location of the occluded area.

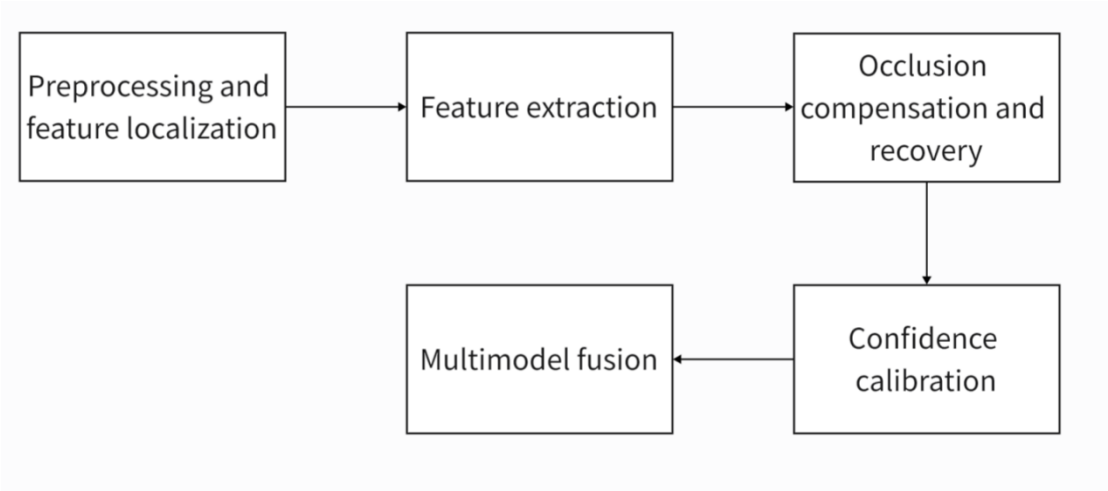

Figure 1: Process of OFR (picture credit: original)

2.2.2. Introduction of CNN

CNN is a deep learning model designed for processing grid-like data, such as images and videos. The core idea is to achieve efficient feature extraction using local receptive fields, weight sharing, and spatial down-sampling. As shown in Figure 2, CNN comprises multiple layers, including convolutional, pooling, and fully connected layers. The convolutional layer is responsible for extracting features, and the pooling layer is used to reduce the data dimension, thereby reducing computational complexity and preventing overfitting. An essential feature of CNN is local connectivity; each neuron is only connected to a part of the input data, which can effectively capture spatial features. At the same time, weight sharing means that the same convolution kernel slides across the entire input image, thereby reducing model parameters.

CNN performs well in areas such as image classification, object recognition, and facial recognition. It can automatically extract features from face images, including facial details and texture information.

DOI: 10.54254/2755-2721/155/2025.GL23384

When dealing with the occlusion problem of face recognition, the base CNN can directly learn the mapping from raw pixels to identity embedding based on large-scale face datasets (such as MS-Celeb-1M mentioned above). When combined with the attention mechanism, it can identify areas of the face that are obscured and focus on features that are not occluded, thus improving expression recognition accuracy [8]. This approach not only enhances the robustness of the model but also improves the adaptability to complex scenarios.

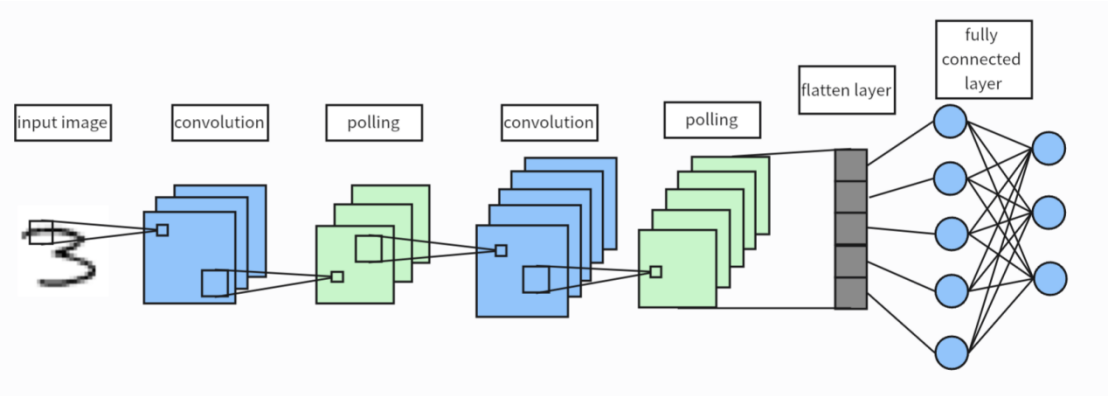

Figure 2: Structure of CNN (picture credit: original)

2.2.3. Introduction of VAE

VAE is a generative deep learning framework based on probabilistic graph models. The core idea is to learn the hidden variable distribution of input data through Variational Inference and generate new samples based on this distribution. The structure of a VAE, shown in Figure 3, is similar to that of a traditional autoencoder. Still, the encoder has two parts: one for calculating the mean of the underlying variable and the other for estimating the variance, which allows the VAE to map the input data onto a probability distribution rather than just a fixed point. Additionally, VAE can generate new data samples based on variations of its training data. This ability makes VAE widely used in image generation and data enhancement. VAE generates new data by compressing input data into a low-dimensional latent space and sampling it in that space. This method can effectively deal with directional probability models with continuous latent variables.

VAE can generate realistic face images, significant for data enhancement and privacy protection. By training VAE models, potential representations of face images can be learned, and new photos similar to the training set can be generated. This ability makes VAE especially important in application scenarios that require large amounts of face data. For example, when training a face recognition system, developing new face images can expand the data set in the face image compression and denoising processing. By learning the potential representation of a face image, VAE can compress image data efficiently. At the same time, the noise is removed during the reconstruction process to improve the image quality. VAE can remove bias from the dataset by learning potential variables. This allows rare features, such as faces wearing glasses or hats, to get more attention in training so that the model can improve the recognition accuracy of these features.

DOI: 10.54254/2755-2721/155/2025.GL23384

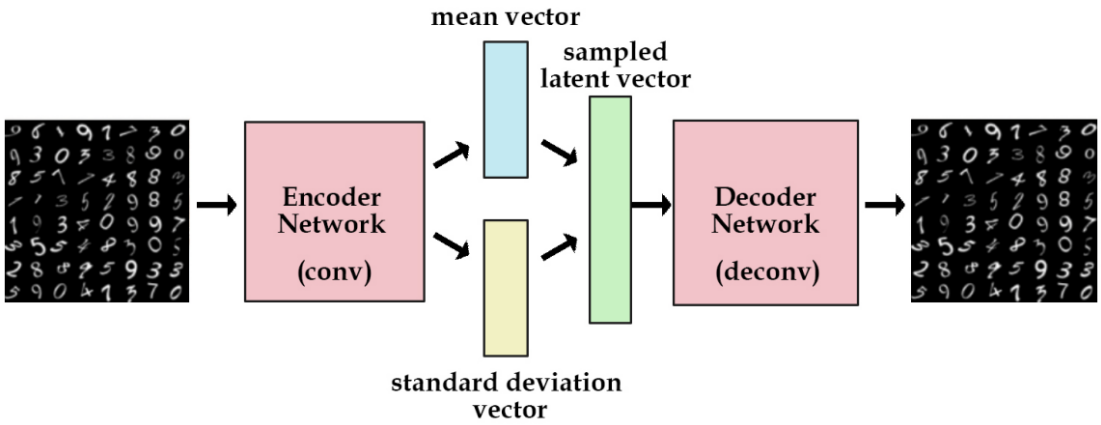

Figure 3: Framework of VAE (picture credit: original)

2.2.4. Introduction of GAN

GAN is a generative deep learning framework whose core idea is to realise data distribution modelling through the adversarial training of the Generator and Discriminator. The Generator is responsible for generating fake data, learning to create real-looking data samples from random noise. The Discriminator is responsible for distinguishing between real data and generated data, and its goal is to identify whether the input data is a real sample correctly. Generators and discriminators are trained against each other. Generators try to generate more realistic data, while discriminators constantly improve their ability to identify forged data. This adversarial process pushes the generator to continuously enhance the samples it generates, bringing them closer and closer to the characteristics of the real data. The process is shown in Figure 4. GAN is widely used in many fields, such as image generation, video generation, image restoration and style transfer. They can be used to create works of art, synthesize high-quality images, enhance data sets, and more.

In the face recognition field, GAN can produce high-quality face images, especially when processing low-quality or blurry photos. Through the adversarial training of the generator and discriminator, GAN can effectively repair and enhance the details and quality of face images. This technique is fundamental in cross-domain face recognition and can help solve recognition problems in different environments. When dealing with occlusion problems, GAN can restore occluded face images. Different head poses are simulated to improve the accuracy of the recognition algorithm. It can also be used to generate adversarial samples to test and enhance the robustness of face recognition systems.

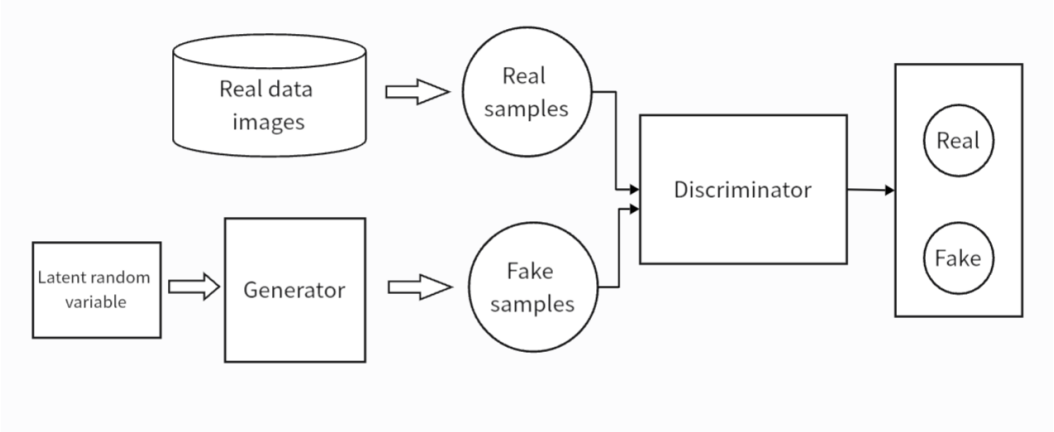

Figure 4: Architecture of GAN (picture credit: original)

DOI: 10.54254/2755-2721/155/2025.GL23384

3. Results and discussion

3.1. Results analysis

As shown in Table 1, VAE and GAN differ significantly in structure, training methods, and the nature of the generated samples. VAE uses a likelihood-based training method, which is usually a more stable training process. By optimizing reconstruction loss and KL divergence to learn the underlying distribution of the data, the statistical properties of the data can be better captured. GAN uses adversarial training, whose process is unstable and prone to mode collapse (mode collapse); the generator may only generate a small number of sample types, resulting in insufficient diversity. Regarding the nature of the sample generated, samples generated by VAE can be fuzzy, especially when reconstructing complex data. Still, they perform well in interpolating potential Spaces and sample diversity. The output of VAE is easier to interpret because its potential space is structured, allowing for better control of the generation process. GAN-generated samples are generally of high quality and precise detail, making them suitable for high-fidelity applications, such as image generation and art creation. However, due to the instability of training, GAN may, in some cases, produce fuzzy or inconsistent samples.

Table 1: Comparison of VAE and GAN

	Construction	Training method	Property of the generated sample
VAE	Encoder and decoder	Likelihood based training	Blur
GAN	Generator and discriminator	Adversarial training	High quality and clear detail

3.2. Discussion

The combination of GAN and VAE (such as VAE-GAN) may become a trend. This hybrid model can leverage the generative capabilities of GAN and the feature learning capabilities of VAE, enabling greater accuracy and robustness in facial recognition. With the development of self-supervised learning technology, GAN and VAE may be used to learn from unlabeled data, further improving the performance of facial recognition systems. This method can effectively train the model in the absence of labelled data.

When dealing with occlusion, the hybrid model also presents many advantages. This model produces higher-quality images: GAN can produce high-quality and high-resolution images. This is particularly important for occlusion processing, where the generated image needs to remain detailed and realistic for subsequent recognition and analysis. VAE is superior in terms of continuity and stability of the underlying space, although it may be blurred when generating images. By combining the characteristics of VAE, the hybrid model can better retain the details and structural information of the image during the generation process, thus improving the quality of the generated image. This model has a stable training process: The training process of VAE is generally more stable than that of GAN and can effectively avoid the typical pattern collapse problem of GAN. By combining the stability of VAE with the generative power of GAN, the hybrid model can maintain high stability during training while generating diverse images. Moreover, VAE learns the distribution of potential Spaces to make the resulting images more feature-diverse. This property enables the hybrid model to create multiple possible occlusion cases when dealing with occlusion, thus improving the robustness and adaptability of the model. This model has better feature extraction and reconstruction capabilities: VAE can effectively extract the features of the input image in the coding phase, which is very important for occlusion processing. By removing the features of the occluded area, the hybrid model can better understand and reconstruct the occluded part. Combining the generative capabilities of

DOI: 10.54254/2755-2721/155/2025.GL23384

GAN and the reconstruction capabilities of VAE, the hybrid model can produce more realistic and coherent images under occlusion conditions. This reconstruction capability enables the model to generate more realistic images when dealing with complex occlusion situations. The model can be applied across modes: hybrid models can be helpful for cross-modal tasks, such as combining images with text information to produce richer content. This capability enables the model to take advantage of additional information to improve the accuracy and reliability of the generation when dealing with occlusion.

4. Conclusion

This study introduces the concept of OFR and highlights the whole process. This paper lists some data sets commonly used in face recognition and presents in detail three models widely used in occlusion-based face recognition: CNN, VAE and GAN. CNN is mainly used for feature extraction. It can effectively improve the accuracy of the recognition of blocked images. VAE and GAN are two commonly used image generation models in face recognition. This paper compares their differences in structure, training methods, and features of generated images in detail. Despite their respective limitations, both models perform well in specific areas. Finally, a hybrid model based on these two models is proposed. Models that combine VAE and GAN produce higher-quality images, and the instability of the GAN training process is solved at the same time. When dealing with occlusion problems, the hybrid model can also generate multiple occlusion types of images. In the future, applying the hybrid model in practical scenarios will be considered the research objective for the next stage. The research will use the VAE-GAN model to OFR scenarios in daily life, such as public safety monitoring, smart homes, etc, then evaluate its performance in a real-world environment and optimize accordingly.

References

[1] *Ngugi, L. C., Abelwahab, M., Abo-Zahhad, M. (2021). Recent advances in image processing techniques for automated leaf pest and disease recognition–A review. Information processing in agriculture, 8(1), 27-51.*

[2] *Li, M., Huang, B., Tian, G. (2022). A comprehensive survey on 3D face recognition methods. Engineering Applications of Artificial Intelligence, 2022, 110: 104669.*

[3] *Dan, Z., Raymond, V., Luuk, S. (2021). A survey of face recognition techniques under occlusion. IET Biometrics, 10(6), 581-606.*

[4] *Wang, X., Guo, H., Hu, S., et al. (2023). Gan-generated faces detection: A survey and new perspectives. ECAI 2023, 2533-2542.*

[5] *Liu, Y., Luo, G., et al. (2024). Adaptive Face Recognition for Multi-Type Occlusions. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 1.*

[6] *Huang, F., Tang, X., Li, C., et al. (2024). Cyclic style generative adversarial network for near infrared and visible light face recognition. Applied Soft Computing, 150, 111096.*

[7] *John, S., Danti, A. (2024). Lightweight Model for Occlusion Removal from Face Images. Annals of Emerging Technologies in Computing (AETiC), 8(2), 1-14.*

[8] *Li, Y., et al. (2019). Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism. IEEE Transactions on Image Processing, 28(5), 2439-2450.*